LLM Apps

Cost of LLM Apps

Types of cost

- Inference cost.
- In-context learning cost.
- Training cost.

Inference Cost

- You pay for the number of tokens processed.

- This cost can escalate very quickly.
  - Summarizing a page in ChatGPT uses 700 tokens and costs 0.015 dollars.
  - Summarizing 1500 pages uses 1M tokens and costs 20 dollars.

In-context learning cost

- The most common form of in-context learning is to build an LLM App using the RAG technique:
  - Use a foundational model to generate language.
  - Add a private database.

- It is a cheaper alternative to pre-training or fine-tuning.
- It has no training costs.
- It is the most recommended approach for most projects. Building an LLM from scratch and fine-tuning are very expensive alternatives.

Fine tuning

- Fine-tuning: training a foundational model with your own database. It has a prohibitive cost in consumption and GPUs for the vast majority of companies.

Training a model from scratch

- Training from scratch: training an LLM model from scratch requires hundreds of thousands of computation hours. The cost increases exponentially with the size of the model and the training database. It has a prohibitive cost in consumption and GPUs for the vast majority of companies.